Some objects are more equal than others: measuring and predicting importance

Merrielle Spain and Pietro Perona

California Institute of Technology
{spain, perona}@caltech.edu

Abstract. We observe that everyday images contain dozens of objects, and that humans, in describing these images, give different priority to these objects. We argue that a goal of visual recognition is, therefore, not only to detect and classify objects but also to associate with each a level of priority which we call 'importance'. We propose a definition of importance and show how this may be estimated reliably from data harvested from human observers. We conclude by showing that a first-order estimate of importance may be computed from a number of simple image region measurements and does not require access to image meaning.

1 Introduction

'Image understanding', the grand goal of machine vision, is about computing meaningful and informative semantic descriptions from images.

Progress in visual recognition has been breathtaking during the past 10 years. We now have algorithms that can recognize individual objects accurately and quickly [1], classify scenes [2], and learn new categories with little supervision [3–7].

What are the next steps toward image understanding? A full description of complex scenes, currently appears to be out of reach (although there is interesting work in that direction [8]). An intermediate goal is generating a list of keywords for each picture. This simpler description would be useful for indexing into large image databases (think of flickr.com's keyword system) and it would be readily understandable by humans. How should such a list be produced? As we shall see later, medium-resolution images of everyday scenes contain dozens of recognizable objects. A number of research groups are making quick progress on simultaneous recognition and segmentation [9–11]. However, rattling off an alphabetized list of nouns would not be particularly informative — not all objects are equal. So our goal is to produce a list of the *important* objects in the scene. We formalize the concept of importance as

> An object's *importance* in a particular image is the probability that it will be mentioned first by a viewer.

This paper is about defining, measuring, and predicting the importance of objects in images. Figure 1 depicts how our ideas fit together. Section 2 describes how we collect words from human observers. In Section 3 we explore how

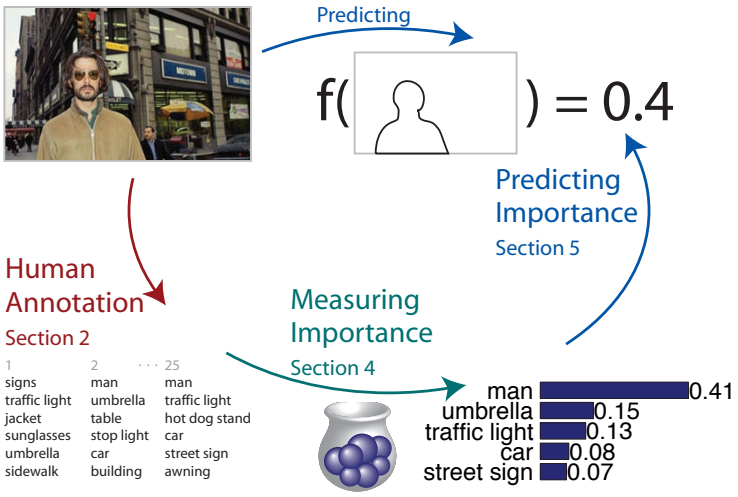

Fig. 1. Which objects matter in a scene? We can measure the importance of an object in a photo by combining lists of the objects named by different viewers (Section 2). To this end, we introduce an urn model, treating object naming as drawing balls from an urn (Section 4). Using these measurements we learn a function to predict object importance directly from photo regions (Section 5).

many objects there are in individual images and collections of images. Section 4 introduces a model for object naming based on importance. We show that this model accounts for both object naming frequency and order. This model, in turn, suggests a method for estimating importance from lists of objects produced by human observers. Section 5 explores whether object importance may by predicted directly from bottom-up visual properties of an object. We conclude in Section 6 with a discussion of our main findings.

2 Human Annotation

Intuitively, an important object is one that could help you identify or recreate the scene. In this section we describe how we collect data that enables measurement of importance.

2.1 Previous Work

The ESP game, by Ahn & Dabbish [12], presents the same image to two players who cannot communicate. Their task is to produce the same word in as few tries as possible. When the players produce the same word, the game ends, banning that word for future plays involving the same image and different players. It is intuitive that the word must be, in some way, related to the image. When

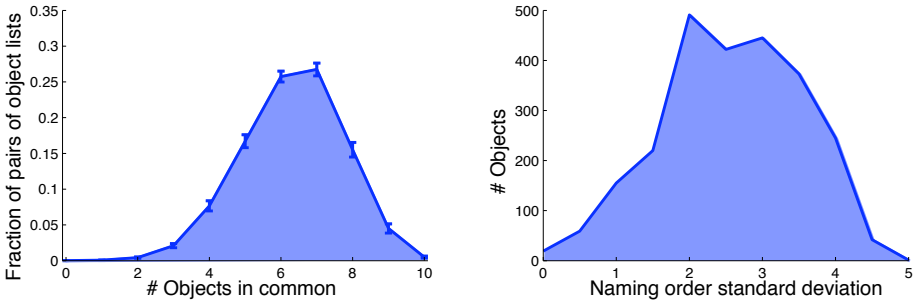

Fig. 2. Humans generate different lists of objects. When two humans independently name 10 objects in a single image, only two thirds of their lists tend to overlap *(left)*. It is extremely rare for the lists to contain all the same objects. The standard deviation in the order that humans name an object in a particular image is large *(right)*.

multiple games are played on the same image, a list of words is produced, one per game. The order of words in the list could measure importance. However, there are weakness in this approach: words are not always object names (e.g. funny) and word order is a noisy correlate of importance since only a pair of players need to produce a given word.

In LabelMe [13] users name objects and outline their contours with mouse clicks. A user may annotate as many objects as they like in an image. Results from previous users are visible to following users, so each object instance is annotated at most once. Elazary and Itti consider the object annotation order a measure of *interestingness* [14]; however, as partial annotations are passed on to new users, a single ordering is produced. Figure 2 shows that when you compare lists of objects named by several humans for the same photo, different objects are named and the object naming order varies wildly. Hence a single list will not capture the statistics of object naming.

2.2 Our Approach

We needed an unbiased, representative, and meaningful set of scenes for our experiments. By 'unbiased', we mean that that the choice of scenes should be as independent as possible from the experimenters and the purpose of the experiment. By 'representative', we mean that the collection of images should sample human visual experience as broadly as possible. However, if the images had been collected completely at random, that is by attaching the camera to a someone's head and snapping one picture per minute for a day [15], most pictures would turn out to be uninterpretable and irrelevant; by 'meaningful', we mean that the images should represent meaningful moments in a person's visual experience. We selected our gallery of 97 pictures from Stephen Shore's collections 'American Surfaces' and 'Uncommon Places' [16, 17]. Shore took these pictures while

traveling in North America in the 70's and 80's and were meant to be a visual diary of his experience. Figures 3 and 4 show sample images.

Through the Amazon Mechanical Turk, observers (English speakers in the U.S.) named 10 objects that they saw in a scene photograph. Each photograph was displayed with a 600 pixel diagonal and annotated by 25 different observers. While previous approaches produce a single word list, we have 25 ordered lists for each image–hence we can use statistical regularities to quantify the importance, not just order of objects.

We used WordNet [18] primary definitions to map synonyms and plurals to the same word, and match missed synonyms (such as misspellings) by hand. As Figure 2 shows, the objects that a viewer names and the order in which a viewer names them vary wildly. Figures 3 and 4 show median order vs. naming frequency (across viewers) for our sample images. Each point corresponds to an object; if an object is mentioned by 35% of the viewers, it has a .35 x-coordinate. The y-coordinate represents the median naming order of the object. So if three observers name a particular object, and it is named 1st, 4th, and 10th, then the y-coordinate is 4. Section 4 introduces a simple model, the urn model, that generates object sequences from object importances, enabling us to measure object importances from human generated sequences.

3 Object counts

A first observation we make in assessing our data, is that there are many objects named in each picture (see Figure 5). For each image, some of these objects are mentioned by a few observers, while other objects are mentioned by many observers. The heights of the curves in Figure 5 quantify how rich with objects images, environments, and the world (full collection) are. The dashed line shows that many fewer objects are named by at least 5 people. The difference between the solid and dashed lines is the number of objects that are rarely named, in that 80% of people don't name them. Hence Figure 5 shows that there are many recognizable objects in a scene photograph, but few objects are preferred by viewers. Hence we need to quantify how important an object is, in order to describe a scene meaningfully with the objects that it contains.

4 Measuring Importance

As proposed in the introduction, we define an object's *importance* in a photo as the probability that a human observer naming objects will name it first. In principle, we would need an extraordinary number of observers to be able to directly calculate the importance of all the objects in a picture: some objects' importances may be less than 1%, and we would need hundreds of observers to determine that. In this section we show that it is possible to measure an objects' importance from relatively few observers if we are willing to model the process that generates an observer's sequence.

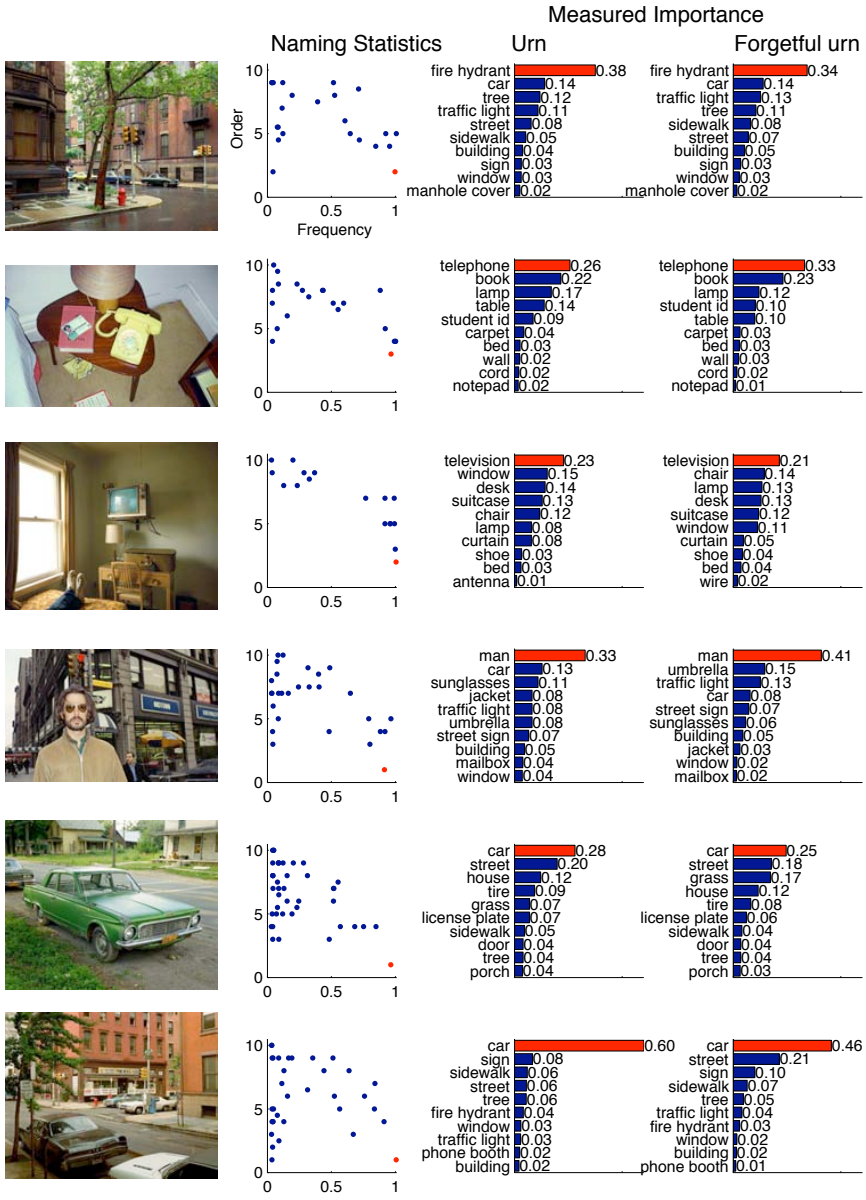

Fig. 3. Examples where the urn model fits the data well. An object's *(dot's)* median order named and frequency mentioned *(2nd column)* are in agreement given the model. In these cases, the forgetful urn *(4th column)* produces similar measured importance to the urn model *(3rd column)*.

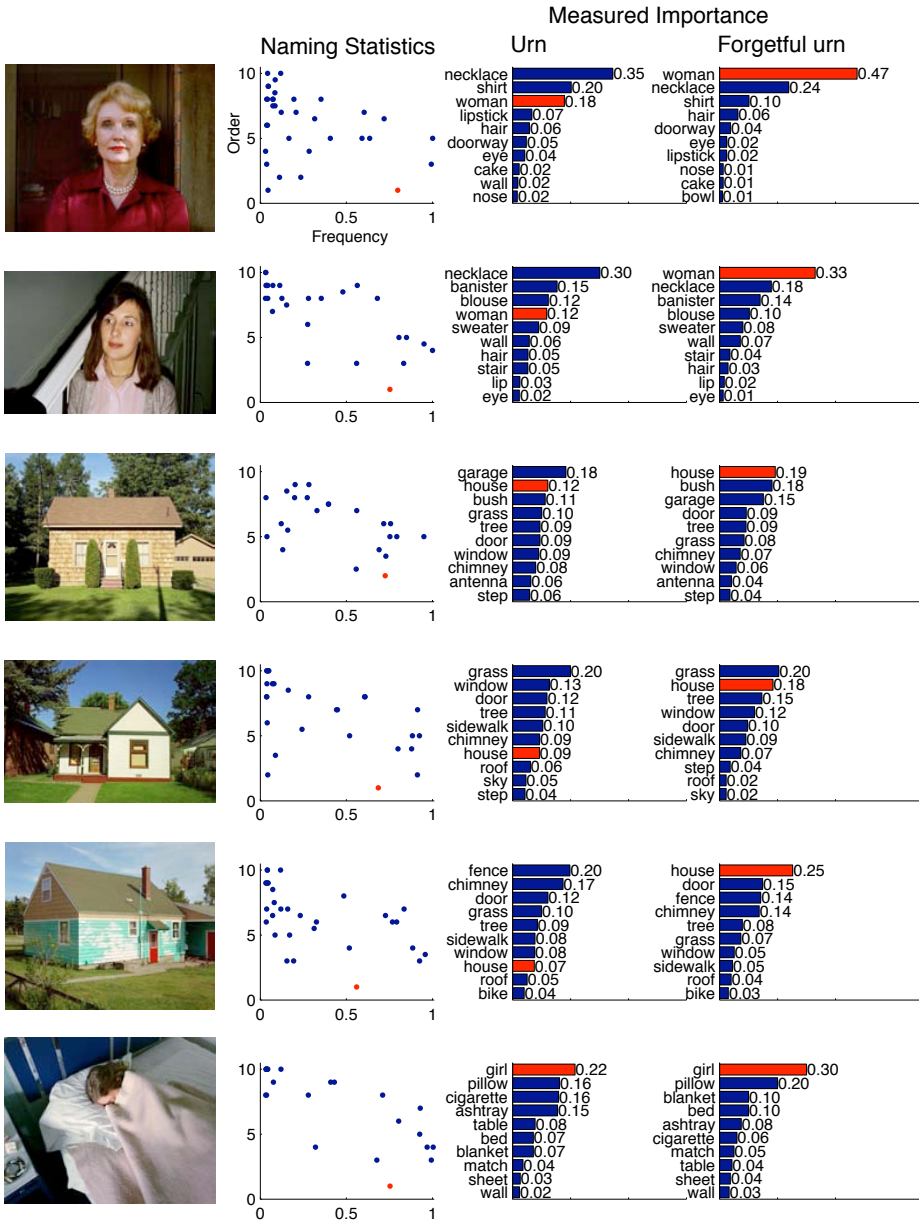

Fig. 4. In many photos with a strong central object, the urn model fails to identify the most important object. The poor behavior of the urn model *(3rd column)* is due to the fact that some viewers fail to name the central object *(red dot)*, while the viewers that name it, name it early *(2nd column)*. We propose a modified model, the forgetful urn *(4th column)* to solve this problem.

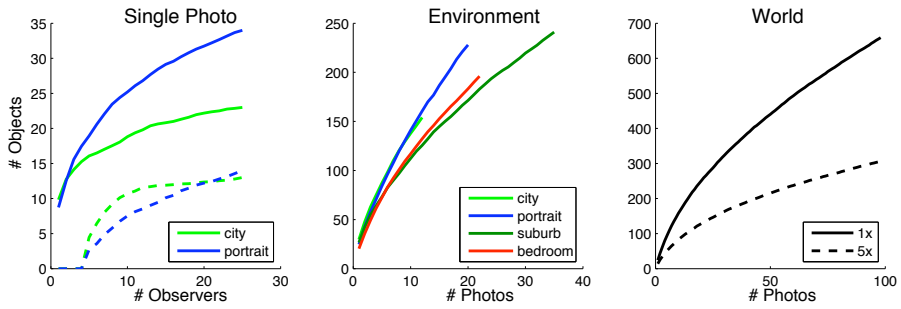

Fig. 5. The number of unique objects named in a photo or collection of photos is surprisingly large and most of these objects are rarely mentioned. The number of unique objects *(solid lines)* named in a photo (top in Figures 3 and 4) as observers are added *(left)*, in an environment as photos are added *(middle)*, and in the full collection as photos are added *(right)*. Many fewer objects have been named by at least 5 viewers *(dashed lines)*. This *(height of the solid line)* displays the sheer number of objects and the large proportion *(solid line - dashed line)* of rarely named objects.

4.1 Urn model

We model the process of naming objects in an image with the process of drawing balls from an urn without replacement (see Figure 6). The balls are different sizes, affecting their probability of being chosen first. Thus, a ball's size represents the importance of the corresponding object. We represent multiple players by repeatedly refilling the urn with the same set of balls and selecting new sequences independently of each other.

Figure 7 shows that randomly drawn lists (of balls) from the urn model can reproduce, at least qualitatively, the order and frequency found in human-generated lists of object names. The synthetic data does not actually correspond to objects in images, but to abstract balls in the urn model. This is a phenomenological model with matches our observations for many images. In order to estimate the size of the balls (the importance of each object) from our human data, we follow the inverse process, that is starting from lists of objects one computes the Maximum Likelihood (ML) values of the parameters of the model (Figures 3 and 4). However, we note that for many photographs composed with a central object, 10-30% of viewers fail to mention that object, (see Figure 4). In the next section we describe our ML method and provide a solution to this problem.

4.2 Fitting the Urn

In order to estimate importance, we could count how often an object is named first, however, this squanders naming order information. When we have limited data (human annotations), finding the Maximum Likelihood Estimator (MLE) of urn model parameters improves upon the direct calculation of importance.

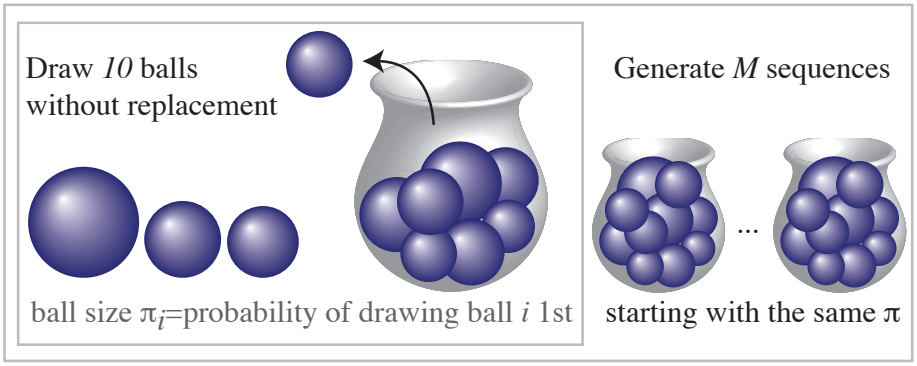

Fig. 6. Urn model relates object importance to named sequences of objects. An urn (image) is filled with balls (objects), having probabilities π_i of being drawn first (importances). 10 balls are drawn (named) from the urn without replacement, creating a sequence. M sequences are drawn.

This is a special case of estimating the weights from a Multivariate Wallenius' Noncentral Hypergeometric Distribution, which requires numerical methods [19]. Manly showed that the weights can be estimated if there are many (>10) balls with the same label being drawn [20], but we have only one ball with each label, so we cannot use his approach. To measure importance via the urn model we need to calculate the probability of observing a set of sequences given the object importances π_i.

Each sequence consists of 10 balls w_i^m (ith ball in the mth sequence) drawn independently without replacement (out of N balls, where $N >> 10$), so the probability of drawing a particular sequence of balls $(w_1^m, ...w_{10}^m)$ is

$$\prod_{n=1}^{10} p(w_n^m | w_{n-1}^m, ...w_1^m) \; . \tag{1}$$

However, we are drawing balls without replacement, so this equation is subject to the condition $w_i^m = w_j^m \implies i = j$. When we draw the nth ball of a sequence, $n-1$ balls have been removed from the urn. The probability that the ball labeled w_n^m is the nth ball drawn is therefore

$$p(w_n^m | w_{n-1}^m, ...w_1^m) = \begin{cases} 0 & \text{if } \exists i \in [1, n-1] : w_i^m = w_n^m, \\ \frac{\pi_{w_n^m}}{1 - \sum_{i=1}^{n-1} \pi_{w_i^m}} & \text{otherwise,} \end{cases} \tag{2}$$

where π_i is the probability that ball i is drawn first (from a fresh urn) and $\sum_i \pi_i = 1$. However, as Figure 4 shows, sometimes viewers fail to mention the most meaningful object. We believe that, when an object is very obvious, some subjects may quickly move beyond it without mention. We treat this as the first

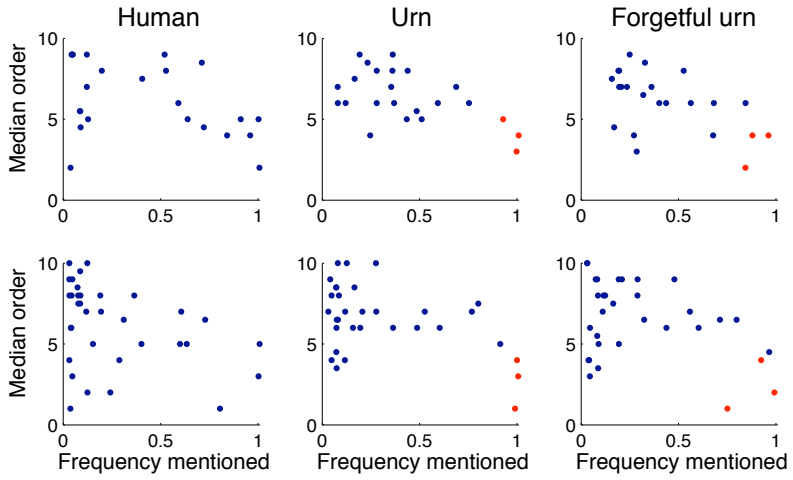

Fig. 7. The urn model reproduces peculiar characteristics of an object's *(dot's)* median order vs. frequency mentioned. Humans *(left)* and the urn model *(middle & right)* produce similar plots. The forgetful urn reproduces how, in some photos, viewers name the most important object either early or never, as if sometimes discarding the 1st ball they drew *(bottom right)*. In synthetic data, the 3 most important objects *(red dots)* are named early and often.

ball being drawn and discarded, and hence excluded from the sequence.[1] We deal with this strange problem in a non-standard way. Consider a sequence of balls where the first ball has been discarded (i.e. really drawn 1st, but not listed); the ball is most likely $\text{argmax}_{j:\forall_i j \neq w_i^m} \pi_j$, that is the most probable of the balls which haven't been mentioned. In this case π_j will likely be large. For a sequence of 10 balls in which the first ball was not dropped, π_j will probably be small. Hence, if we include $\max_{\forall_i j \neq w_i^m} \pi_j$ in the normalization, we obtain estimates of π_i that are not far from the correct ones when the first ball is not dropped and most often the correct estimates when the first ball is dropped. This changes Equation 2 to

$$p(w_n^m | w_{n-1}^m, ...w_1) = \frac{\pi_{w_n^m}}{1 - \max_{\forall_i j \neq w_i^m} \pi_j - \sum_{i=1}^{n-1} \pi_{w_i^m}} \,. \tag{3}$$

Since we have M independent sequences, the likelihood of our observation is

$$p(obs) = \prod_{m=1}^{M} \prod_{n=1}^{10} \frac{\pi_{w_n^m}}{1 - \max_{\forall_i j \neq w_i^m} \pi_j - \sum_{i=1}^{n-1} \pi_{w_i^m}} \,. \tag{4}$$

To measure importance $\pi_{w_i^m}$, we maximize the log-likelihood $log(p(obs))$,

[1] So the rigorous definition of importance is the probability that a ball is drawn first, regardless of whether it is discarded.

$$\sum_{m=1}^{M}\sum_{n=1}^{10}\log \pi_{w_n^m} - log(1 - \max_{\forall_i j \neq w_i^m} \pi_j - \sum_{i=1}^{n-1} \pi_{w_i^m}) \; . \tag{5}$$

Using synthetic data generated by an urn model Monte Carlo with different parameter settings, we see that the MLE of Equation 5 enables us to estimate the parameters more precisely than direct calculation from the observed frequency. Particularly, Figure 8 shows that the urn model and the forgetful urn model (the model containing the $\max_{\forall_i j \neq w_i^m} \pi_j$ term) are much closer to the true importance distribution in the K-L Divergence, $D_{KL}(\pi||\hat{\pi})$ [21]. The forgetful urn outperforms the original urn model when the first ball has a nonzero probability of being dropped, and is equivalent when the first ball isn't dropped. These are the mean K-L Divergence values over 5 importance distributions, each with 10 sets of 25 sequences drawn, shown at 4 probabilities of dropping the first ball.

Figures 3 and 4 display the importances of the 10 most important objects in our sample images, using the urn MLE and forgetful urn MLE.

One could wonder if our definition of importance captures objects that may never be named first. For instance in a photo of Batman and Robin, Robin may never be named first, yet he is important. In this example Robin's consistently second position violates the independence assumption of our model. The fitting will then assign a high importance to Robin. Empirically, we can take data from the urn model and move the second most important ball to second place every time it is drawn first. In our simulations this change does not significantly decrease the estimated importance of this ball (Wilcoxon Rank Sum Test).

Optimization Note: There are as many parameters π_i as objects mentioned. Figure 5 shows that this number can get large, which results in poor convergence. However if we limit our optimization to the 10 most frequently named objects and set the others to a small constant, our convergence using fmincon (repeatedly and perturbing the solutions) in the Matlab Optimization Toolbox is quite good.

5 Predicting Importance

Is it possible to automatically predict importance from image measurements without using human observers? A number of researchers are working on the problem of segmentation and recognition. In this paper we wish to focus on the orthogonal problem of estimating importance once the objects have been detected and segmented. Since no such system works sufficiently well nowadays, we segment the image by hand.

Using a training set of 30 images (687 objects), a validation set of 10 images (217 objects), and a test set of 35 images (800 objects), we fit a generalized linear model $log(\pi_i) = \sum_j X_{ij} b_j$. Here π_i is the measured importance of object i (from Section 4) in training and the predicted importance in testing. Our features consider the composition of the photo in relation to an object's outline; X_{ij} is object i's value for feature type j and b_j is the weight of the jth feature.

As there were many more unimportant objects than important ones (importance > 0.07), we selected an equal number of each, using outlier removal. Our

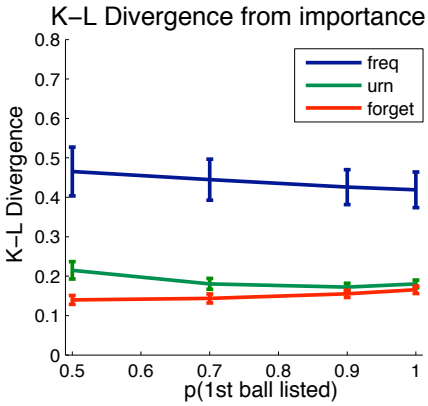

Fig. 8. The urn models measures importance more precisely than directly calculating how often an object is named first. An urn model Monte Carlo simulation generated sequences (with different probabilities of listing the first ball), and we used these sequences to estimate the synthetic parameters. The urn models' MLEs *(green, red)* are closer to the true parameters than the first mentioned frequency estimate *(blue)*. The forgetful urn model *(red)* is better than the plain urn model *(green)* when the first ball could be discarded, and equivalent otherwise.

outlier removal was inspired by Angelova et. al's [22] idea that noisy data should be predicted poorly by a partial model that it didn't influence. We built many partial models and predicted the low importance training objects with them; we discarded the objects that gave the largest squared error across partial models.

We added features to each regression greedily, by choosing the feature that most reduced training error. We grew a regression starting at each feature, stopping growth at the lowest validation error. Finally, we selected the regression with the lowest training error. Out of a list of 30 features, the chosen features (in order of choice) were: distance from the image center to the closest part of the object, minimum distance from the object to the 4 points that divide the image into thirds (taken from the thirds rule of photographic composition), minimum vertical distance above the image center, maximum saliency[23] on the object, mean number of overlapping objects, total saliency over object, total blurred saliency over object (2 pixel blur on saliency map), maximum vertical distance below the image center, and maximum horizontal distance from the image center.

Our regression predicted importance (from photo patches) which we compared to measured importance (from human data in Section 4). We evalated the quality of such predictions quantitatively by building a binary classifier: objects were classified as as having higher or lower importance than a given importance threshold. Figure 9 shows the ROC curves at 3 importance levels, illustrating that our prediction method can discriminate very important objects. If we want

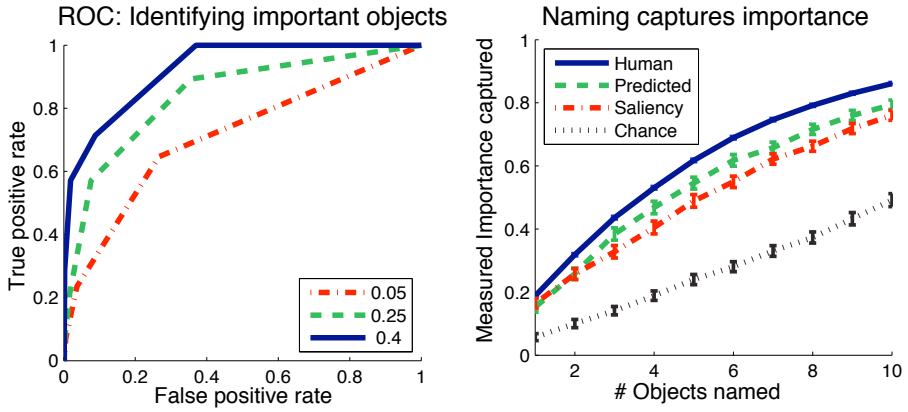

Fig. 9. Regression on image region properties can predict object importance. We define 'important' objects as having a measured importance > 0.05 *(red)*, 0.15 *(green)*, 0.4 *(blue)* and produce an ROC Curve for each definition *(left)*. We calculate how much measured importance has been captured by the first n objects named *(right)*. Humans *(blue)* outperform predicted importance *(green)*, which in turn outperforms the total saliency of an object *(red)*, which is much better than chance *(black)*.

to name objects to capture as much measured importance as possible, we can sum the measured importance of the first n objects named. The naming order is obtained by sorting objects by predicted importance (or total saliency). Figure 9 shows the performance of predicted importance is sandwiched between human naming and saliency; it also shows that all 3 greatly outperform chance.

For a more intuitive evaluation, Figure 10 shows the importance predictions for sample photographs, the most noticeable discrepancy between automatically predicted importance (Figure 10) and importance estimated directly from human subjects (Figures 3 and 4) is that the fire hydrant was more important to people than our regression predicted.

6 Discussion

We asked a number of human observers to list the objects that they saw in images of complex everyday scenes; each image (of 97) was annotated by 25 observers. The data we collected shows that our visual world is rich (Figure 5): there are dozens of things that may be recognized in each image.

A number of research groups are making progress on simultaneous recognition and segmentation. Here we study the complementary problem of importance. Not all objects are equal: in a given image some are mentioned by all observers and some by few, some are mentioned early and some later. This suggests that we should not be content for our recognition algorithms to return a laundry list

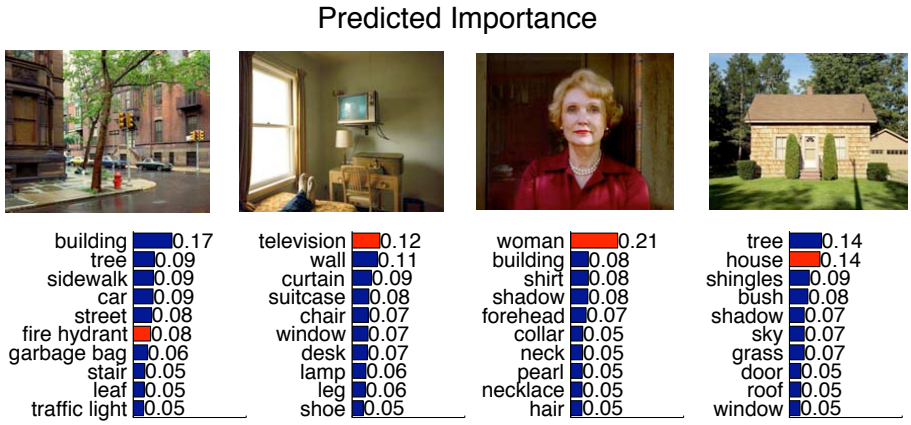

Fig. 10. Examples of predicted importances for sample images. The object with the highest measured importance *(red)* tends to have a high predicted importance. We train on image region properties with measured importance (Figures 3 and 4) as the desired output. Our testing output is predicted importance.

of things that are present in the image. We suggest that a complete system will require both recognition-segmentation and importance.

We defined 'importance' as the probability that an object is mentioned first by a human observer. We provided a methodology for measuring importance of objects in images from the data provided by our human observers. We noticed that human observers sometimes miss the most obvious object. We proposed a procedure to measure importance that takes this phenomenon into account. We found experimentally that our measurements of importance are reliable on synthetic data (Figure 8) and intuitive on human data (Figures 3 and 4).

One could worry that it may be impossible to assess an object's 'importance' automatically until the meaning of an image is understood [24]. We explored how far can one go in estimating object importance automatically from bottom-up image measurements. We found that a small number of simple measurements go a long way towards predicting importance (Figure 9).

Finally, it should be pointed out that this work is about photographs taken by humans — our findings may not generalize to haphazardly captured photographs.

Acknowledgments This material is based upon work supported under a National Science Foundation Graduate Research Fellowship, National Institute of Mental Health grant T32MH019138, Office of Naval Research grant N00014-06-1-0734, and National Institutes of Health grant R01 DA022777. We would like to thank Antonio Torralba for insightful discussions and Ryan Gomes and Kristin Branson for useful corrections.

References

1. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (2004)
2. Oliva, A., Torralba, A.B.: Scene-centered description from spatial envelope properties. In: Biologically Motivated Computer Vision. (2002) 263–272
3. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: ECCV (1). (2000) 18–32
4. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR (2). (2003) 264–271
5. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: ICCV. (2005) 370–377
6. Grauman, K., Darrell, T.: Efficient image matching with distributions of local invariant features. In: CVPR (2). (2005) 627–634
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2). (2006) 2169–2178
8. Barnard, K., Forsyth, D.A.: Learning the semantics of words and pictures. In: ICCV. (2001) 408–415
9. Russell, B.C., Efros, A.A., Sivic, J., Freeman, W.T., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: Proceedings of CVPR. (2006)
10. Andreetto, M., Zelnik-Manor, L., Perona, P.: Unsupervised learning of categorical segments in image collections. In: Computer Vision and Pattern Recognition (CVPR08). (2008)
11. Todorovic, S., Ahuja, N.: Extracting texels in 2.5d natural textures. In: Proceeddings of ICCV. (2007)
12. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: CHI. (2004) 319–326
13. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. Technical report (2005)
14. Elazary, L., Itti, L.: Interesting objects are visually salient. Journal of Vision **8** (2008) 1–15
15. Mayer, M., Switkes, E.: Spatial frequency taxonomy of the visual environment. Investigative Ophthalmology and Visual Science **26** (1985)
16. Shore, S.: Stephen Shore: American Surfaces. Phaidon Press (2005)
17. Shore, S., Tillman, L., Schmidt-Wulffen, S.: Uncommon Places: The Complete Works. Aperture (2005)
18. : (Wordnet)
19. Fog, A.: Calculation methods for wallenius' noncentral hypergeometric distribution. Communications In statictics, Simulation and Computation **37** (2008) 258–273
20. Manly, B.F.J.: A model for certain types of selection experiments. Biometrics **30** (1974) 281–294
21. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics **22** (1951) 79–86
22. Angelova, A., Matthies, L., Helmick, D.M., Perona, P.: Fast terrain classification using variable-length representation for autonomous navigation. In: CVPR. (2007)
23. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks **19** (2006) 1395–1407
24. Yarbus, A.: Eye movements and vision. Plenum Press, New York (1967)